Evaluating Usability of a Real Time Augmented Reality Audio Captioning System

Salaar Liaqat
University of Toronto
sliaqat@cs.toronto.edu

ABSTRACT

People who are deaf or hard-of-hearing have expressed interest in systems for real-time speech captioning using head-mounted displays. These systems show captions to the user outside of their center of vision. With the captions readily available at a glance, delays are detrimental to the user experience. The extent of acceptable caption delays on head-mounted devices has not been explored. Works that have developed real-time captions have avoided discussion of latency or users reported that the latency was too high.

In this paper we leverage an edge computing architecture to provide real-time captioning on a head-mounted display. We run a study to evaluate the latency requirements of captions and we compare the results to the latency of our system. We run a second study to evaluate the usability of our system by having participants engage in a conversation using a head-mounted device for input. We find that our system meets the latency requirements of study 1, and that participants found that latency acceptable in study 2.

1 INTRODUCTION AND MOTIVATION

According to the Canadian Association of the Deaf [1], approximately 10% of Canadians are hard of hearing and 1% are linguistically deaf. In daily life, people who are hard of hearing or deaf often have trouble understanding speech and thus, take certain approaches to assist their ability to communicate. The most common approaches are a combination of hearing aids, lip reading and social assistance [5]. However, hearing aids and lip reading are both limited by being most effective when the sound originates from in front of the user. In response to these limitations, [5] found that real-time audio captioning was a sought after feature for head-mounted displays.

Wearable devices provide users continuous access to a computational device with little burden. These devices are outfitted with various sensors, displays and wireless connectivity, enabling the devices to provide rich information to the user unobtrusively. However, they are often limited in terms of computational power, battery life and acceptable thermal levels. Furthermore, applications in assistive technologies, and in particular for wearable devices often have a latency requirement in order to provide an acceptable level of usability.

, *CSC2526, Assistive Technology*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In this paper, we develop a system for real-time audio captioning. The system consists of a mobile application running on a head-mounted display. The application is responsible for recording audio, detecting the presence of speech and uploading speech to a remote server. The server component transcribes audio and sends it back to the head-mounted display. We run a study to evaluate the latency requirements of captions. We also benchmark our system and find that the latency of our system is below the annoyance level that we determined in the first study. We run a second study to evaluate the usability of our system by having participants engage in a conversation using a head-mounted device. The participant rely on the head-mounted device for captions and are later asked about their experience. We find that our system meets the latency requirements of study 1, and that participants found the latency acceptable during use in study 2

Our contributions can be summarized as follows:

- Explore the time requirements of captions through a study
- Develop a low-latency system to provide real time captioning on a head-mounted device
- Explore the usability of the system through a study

The rest of the paper is organized as follows. Section 2 discusses the related work in speech captioning. Section 3 discusses the system and implementation. Section 4 is the evaluation of our system. Section 5 discusses future work, and section 6 present our concluding thoughts.

2 BACKGROUND AND RELATED WORK

The usability of head mounted audio processing has been explored in several works. Most works implement a system to provide audio related information through visual means [5, 6]. The works explore the various design considerations when designing an application for head mounted display. We drew inspiration from these works when designing our system. Most papers that study the usability of head mounted devices for providing audio information avoid studying latencies. Existing prototypes either have high latencies caused by the use of the cloud, or minimal latencies achieved by augmenting the system with hard-coded captions. In either case, minimizing latency is not a common problem addressed in these works.

One example of an existing work, SpeechBubbles [6], uses the Microsoft HoloLens headset to provide real time captions for people who are deaf or hard of hearing. The system provides captions in speech bubbles to provide users the source and order of speech in conjunction with the transcriptions. The system uses the Google Speech API to provide transcription and audio source was obtained through the use of multiple speakers. The paper evaluated the usability of speech bubbles as well as various design considerations, however, latency was not a focus of the paper, but according to

the authors, the transcription would fall a few sentences behind the current audio. Inspired by the potential of the SpeechBubbles, we aim to explore the latency issues that weren't discussed in the paper. We avoid using speech bubbles in our system and aim to use a bare-bones UI to minimize distraction and obtrusiveness.

Real time audio captioning has been thoroughly explored for live broadcasting for television. The FCC describes delayed or off-sync captions as "worthless" [3] and for live audio captioning, an effort should be made to keep captions in sync. The Canadian Radio-television and Telecommunications Commission has defined regulations for broadcasting and captioning. They worked with the English-language Closed Captioning Working Group (EN-CCWG) to define the specific requirements for displaying captions as well as the timing constraints associated with captions. They determined that the lag time for captions should never exceed 5 seconds, and 80% of the captions should never exceed 3 seconds [2]. For our system, we target a latency of below 3 seconds, however, we also run a study to evaluate the usability of delayed captions.

3 SYSTEM DESCRIPTION

We implement a system for real-time audio captioning on a head-mounted display. The head mounted display runs a mobile application and is responsible for sending speech to the server. The server resides on the edge, and is responsible for transcribing speech.

The mobile application was developed for the Epson Moverio BT300 Smart Glasses running Android 4.1. The glasses, shown in Figure 1, contain a quad-core Arm processor, 2Gb of ram and a heads up display.

Figure 1: Epson Moverio BT-300

The server component of the system ran on a server named Mel and consisted of a 14 core Intel Xeon E5-2680 with 128Gb of ram.

3.1 Server Application

The basis of the server application uses the Kaldi Library [7], which is toolkit for speech recognition. It provides a set of machine learning models and a large variety of tools for developing and using these models. Our server uses the Kaldi Gstreamer open-source server to interface with the Kaldi models and provide speech recognition. The server utilizes master/worker setup where the master communicates with the client and worker, and the worker processes the audio segments.

Kaldi Gstreamer provides online speech recognition. This means as speech is being recorded and fed into the server, the server makes predictions using the available data. These predictions are sent to the client when available. Once the client runs out of audio to send, the client must send the server an end-of-stream message. This indicates to the server that audio will no longer be sent, so the server is able to stop listening for incoming packets and make a final prediction for the entire audio segment.

The machine learning model used in this system was the Tedlium Language model [8], which was developed for automatic speech recognition. This model is build on a corpus created by the developers using unsupervised machine learning techniques. From the corpus, they developed a speech recognition algorithm using a multi-layered perceptron. This is a lightweight model but achieves a 17.4% word error rate. Although we chose this model for its quick and accurate transcriptions, the Kaldi GStreamer server was designed so any Kaldi model could be plugged into the system and work with minimal effort.

3.2 Glasses Application

When designing the system, inspiration was drawn from Zhang et al's . [9] work, where the authors developed an edge and cloud based system for delivering augmented reality online video game content. The authors found that the latency from the cloud was too high to play virtual reality video games. To minimize latency during gameplay, the authors split events into events that could be completed locally and those that needed information from the cloud. This resulted in most events being completed on the edge which had an acceptable latency for virtual reality gaming. We utilize the same methodology, however, we use the mobile device for a lightweight processing step and only complete computationally expensive tasks on the edge. Specifically in our implementation, we apply this to detecting speech. Instead of uploading all audio recorded to the server and requiring the server to locate speech, the device runs a speech detection algorithm to locate which audio should be uploaded.

The glasses continuously records audio in 100 millisecond segments which gets run through a speech detection algorithm. To reduce computational complexity and increase battery life, a complicated machine learning model was avoided. Instead, we implemented a speech detection algorithm which extracted frequencies from the audio signal and looked for frequencies within human speech range. The algorithm would wait for 500 milliseconds without speech before reporting that no speech was detected. Although the algorithm is simple, we found it was able to reliable detect speech with very few false positives. A wait time of 500 milliseconds was chosen based on developer decision, but could be a point of research in the future.

Once speech is detected, the client establishes a connection with the server by sending the server the details of the audio recording. For our application, we recorded at 16000Hz on a single channel in 16 bit PCM audio. Once the connection is established, we send audio for as long as speech is detected, after which we send the end-of-signal message. Partial hypotheses are received while audio is being sent, and to reduce user perceived latency, we display predictions as soon as they are received.

The UI of the headset was kept simple. The application consisted of a single page with a black background. This corresponds to a transparent screen on the head mounted display and enabled the users to see past the headset display. The page had a single toggle

Figure 2: Idle screen of application

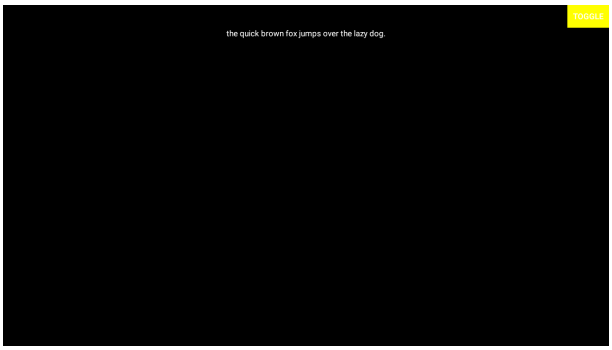

Figure 3: Captions displayed by application

button in the top right page which was red when the application was idle. If the user clicked the button, it would turn yellow and start the speech detection. If speech was detected, the button would turn green and captions would appear on the top of the screen in white text. The button would turn back to yellow if the speech was no longer detected, or red if the button was clicked again. Figure 2 shows the application during the idle stage, and Figure 3 shows the application after a sentence was spoken.

4 INTIAL EVALUATION RESULTS

4.1 Study 1

The first study we ran aimed to determined the delay requirements between speech and their corresponding captions. Although the EN-CCW requires live broadcasting to have a lag time of under 3 seconds for 80% of captions [2], this number is not justified by any trial. To evaluate what lag time was acceptable in a controlled study, we recruited 5 participants, 4 male, aged 22-31. None of the participants where deaf or hard-of-hearing. The participants where shown a 8 minute TV-show clip with captions. The captions started off synced to the audio, however, every 5 captions the delay would increase by 250ms. We informed the participant of this and instructed them to tell us when the captions where noticeably delayed, when they became annoying, and when they became unusable. To increase their dependence on the captions, we played the video clip an a low volume with high levels of background white noise.

Delay	Mean (sec)	Standard Deviation
Noticeable	1.05	0.542
Annoying	2.15	1.194
Unusable	3.35	0.929

Table 1: Study 1 Results

Results from study 1 showed that delays became noticeable after approximately 1.1 seconds, annoying around 2.2 and unusable after 3.3. However, the standard deviation suggests these values can vary significantly depending on the individual. Comparing our results to the requirements of the EN-CCW, if the EN-CCW wants to avoid unusable captions and can settle for annoying in the case for live broadcasting, their requirement of under 3 seconds coincides with our findings. For our system, we want to avoid having annoying levels of delay, so we target of a latency of under 2 seconds.

4.2 System Benchmark

Before evaluating our system through user studies, we benchmark our captioning speed through automated captioning of pre-recorded audio. We setup a client application to run on a computer and send the captions to the server performing the transcription. Since a desktop with wired connection experiences less network latency than a mobile device using WIFI, we simulate network latencies based on the findings of [4] for our benchmarks. To showcase that our system performs better than a cloud based, solution, we simulate the network latency of connecting to the edge and cloud. For measurements, we start timing once the audio is sent to the server, and record when a partial hypothesis is returned, as well as when the final transcription arrives.

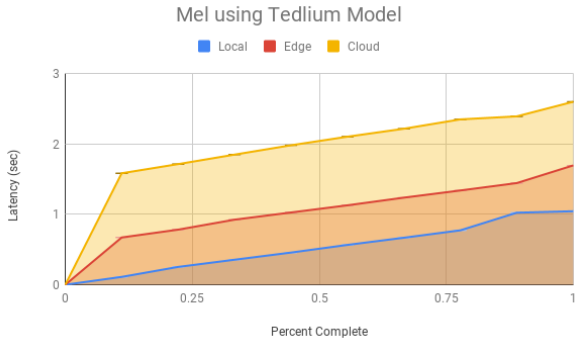

Figure 4: The cost of transcribing audio locally, on the edge (our system) and on the cloud

In figure 4, we can see that the total cost of a transcription on our system (Edge) is approximately 1.8 seconds. With 1 second spent on computing the transcription and .8 seconds spent on sending the transcription. The figure also shows that running the same system on the cloud takes approximately 2.5 seconds, which starts to approach the unusable threshold determined in study 1. The Tedlium model used in our system was lightweight and ran quicker

than other models that were considered. With more computationally expensive models, running on the would result in an unusable system. This benchmark shows that our system is able to perform below the annoyance threshold that we determined in study 1,

4.3 Study 2

In our second study, using the same 5 participants, we asked the participants to wear the head mounted device and to wear a pair of headphones which played white noise. We ensured the participants where unable to hear us intelligibly, and then had a researcher conversed through a set of questions with the participant. If the participant was able to derive the question from the caption, they were instructed to continue with conversation. If the participant didn't understand a question, they could ask the researcher to repeat themselves. After the conversation, the participant where asked for feedback on the delays of the captions, the levels of assistance the captions provided and the overall thoughts on the system.

When asked if there was a delay between when the researcher asked a question and the captions appeared, all participants reported that there was a delay. One participant said the delay was noticeably bad and that it delayed comprehension of the sentence. Three participants said it was noticeable but it was acceptable and didn't slow down the work-flow. The last participant reported that the delay was very good since when they needed to refer to the captions the word they needed was there. Overall, there was positive feedback on the delay.

When asked if the captions assisted them throughout the conversation, all participants reported that the captions were useful. Furthermore, they all reported that if they were in a situation where their hearing level was of that in the experiment, they would resort to such a device. However, participants did have suggestions. All participants responded that the caption size was too small and three participants wanted the captions closer to the center or bottom of the screen. This contributed to a strain on their eyes and they were concerned with prolonged use. Three participants found captions from their own speech annoying but two participants found it didn't bother them. One participant said the partial hypothesis during the speech was annoying and along side this, any incorrect captions were also annoying. They found these aspects of the system broke their train of thought. Three participants said they would have liked the options to see a log of captions and one participant found the head mounted display uncomfortable and too bright.

When participants where asked about any additional smart features they would have liked to see, some reported that they would have liked to see color coded text. One participant said colors should refer to the person speaking in a group environment. Two participants said that the color could correspond to the confidence level of the caption. A general trend in additional features was maximize information aslong as the information is presented intuitively and the UI remained uncluttered.

When asked about any limitations with the system, participants where concerned with noisier environments, such as a checkout in a grocery store or restaurant. Another similar concern was multiple speakers and quick conversations. One participant reported that they may avoid using the head-mounted device simply due to the social stigma around such a device. Furthermore, they said that they found themselves too focused on the display and didn't really pay attention to the speaker. This raised the concern that the head-mounted display may negatively impact social interactions. The last concern was the requirement of having good vision when a high number of deaf or hard-of-hearing individuals are older and also suffer from vision problems.

The overall findings of the system where the system developed was usable and the delay of the captions where not a problem for the most part. However, the system did have some problems with the implementation of the application and participants would like to have some control over customization of the visuals such as size and location of the captions. Participants would also like more features and would like the system to function in noisy environment.

5 LIMITATIONS AND FUTURE WORK

Our studies consisted of five participants, none of whom where drawn from our desired audience for this system. This adds a level of concern in our participatory feedback since the feedback could not be related back to the struggles of those who are deaf or hard-of-hearing. In the future development of this system, having participants that could benefit from the system would provide more insightful feedback.

Our work was focused on the delay requirements of real-time captioning. Because of this, our studies didn't aim to provide the visuals for the best user experience. Based on participant feedback, incorporating features such as controllable caption size and location, caption accuracy confidence, caption history, and longer lasting captions would increase the usability of the system. Although work has been done in providing speech bubbles and audio visualization on head-mounted devices, captioning in an unobtrusive but available-at-a-glance way has yet to be explored.

In a future work of this system, we think it would be better to have two studies using the system rather than the one we had in this paper. The first study would evaluate the design choices and define the best UI with the option of customization where necessary. Then in the second study, evaluate the overall usability, potentially in a long term study. This would provide refined feedback on both the best design and the overall usability.

An interesting problem that arose in study 2 was the device wearer's speech being captioned. Three of the five participants found that distracting. It would be interesting to explore solutions to avoid captioning audio originating from the user. Solutions could exist by using multiple microphones to locate the sound and ignore sounds from the user, or by using a directional microphone. Another approach could be to try and predict when the user is speaking through the accelerometer or a vibration sensor close to the ear hooks. Exploring this options could produce interesting results and could be worthwhile future work.

6 CONCLUSION

In this paper, we look at the usability of real-time captions on head-mounted devices. We focus on the latency issues with captions and head-mounted devices, since both these systems tend to have strict latency requirements. We draw upon literature for designing a system and develop a system that provides real-time captions on

a head-mounted device. We explore the latency requirements for captions by looking at regulations in live TV broadcasts and design a study to verify those numbers. We benchmark our system and show that our system is able to provide captions in a delay that participants reported to be noticeable but not annoying. We also ran a second study to evaluate the overall usability of the system. Our results correspond to our first study and benchmarks, that reported that the captions were noticeably delayed but not annoying. Feedback for improvement was focused on providing better UI and richer information alongside captions. Overall, this paper outlines the latency requirements of real-time captions on head-mounted displays and develops a system which is able to meet those requirements.

REFERENCES

[1] [n. d.]. Statistics on Deaf Canadians. ([n. d.]). http://cad.ca/issues-positions/statistics-on-deaf-canadians/

[2] Canadian Radio-television and Telecommunications Commission. [n. d.]. Broadcasting Regulatory Policy CRTC 2012-362. ([n. d.]). https://crtc.gc.ca/eng/archive/2012/2012-362.htm

[3] Clifford Harrington and Christine Reilly. [n. d.]. Closed Captioning Quality Standards Go Into Effect April 30, 2014. ([n. d.]). https://www.pillsburylaw.com/images/content/6/5/v2/65625/AlertApril2014CommunicationsClosedCaptioningQualityStandardsComp.pdf

[4] Wenlu Hu, Ying Gao, Kiryong Ha, Junjue Wang, Brandon Amos, Zhuo Chen, Padmanabhan Pillai, and Mahadev Satyanarayanan. [n. d.]. Quantifying the Impact of Edge Computing on Mobile Applications. ACM Press, 1–8. https://doi.org/10.1145/2967360.2967369

[5] Dhruv Jain, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E. Froehlich. [n. d.]. Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing. ACM Press, 241–250. https://doi.org/10.1145/2702123.2702393

[6] Yi-Hao Peng, Hsien-Hui Tang, Mike Y. Chen, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, and Yu-An Chen. [n. d.]. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. ACM Press, 1–10. https://doi.org/10.1145/3173574.3173867

[7] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Jan Silovsky, Georg Stemmer, and Karel Vesely. [n. d.]. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (2011). IEEE Signal Processing Society.

[8] Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. 125–129.

[9] Wuyang Zhang, Jiachen Chen, Yanyong Zhang, and Dipankar Raychaudhuri. [n. d.]. Towards efficient edge cloud augmentation for virtual reality MMOGs. ACM Press, 1–14. https://doi.org/10.1145/3132211.3134463